Ανάκληση Πληροφορίας

Διδάσκων –
Δημήτριος Κατσαρός

Διάλεξη 13η: 28/04/2014

Παράμετροι του μοντέλου PageRank

Η παράμετρος α (1/2)

- Η παράμετρος αυτή ελέγχει στην ουσία την προτεραιότητα που δίνεται στη δομή των υπερσυνδέσμων ή στην τηλεμεταφορά
- Είδαμε στην προηγούμενη διαφάνεια ότι οι Brin & Page πρότειναν τιμή .85 για την παράμετρο αυτή
- Γιατί αυτήν την τιμή;
- Ποια είναι η επίδραση του α στο πρόβλημα του PageRank;
- Με α=.5, τότε η επαναληπτική μέθοδος χρειάζεται μόνο 34 επαναλήψεις για να συγκλίνει σε μια ακρίβεια 10^{-10} !!
- Όμως αυτό σημαίνει ότι η τεχνητά εισαχθείσα έννοια της τηλεμεταφοράς θα είναι ίσης σημαντικότητας με τη δομή των υπερσυνδέσμων !?

Η παράμετρος α (2/2)

- Για α=1.0, οι αριθμός των επαναλήψεων για σύγκλιση γίνεται απαγορευτικός
- Ακόμα και για α=.85 απαιτούνται μερικές ημέρες για να επιτευχθεί η σύγκλιση όταν οι πίνακες είναι του μεγέθους του Παγκοσμίου Ιστού
- Απλώς το α=.85 επιτυγχάνει ένα αποδεκτό tradeoff
- Πέρα από αυτό όμως, η παράμετρος ελέγχει και την ευαισθησία του διανύσματος PageRank
- Για τιμές του α κοντά σε 1, τότε ακόμα και μικρές αλλαγές στη δομή του Web Επηρεάζουν σημαντικά τις τιμές PageRank των σελίδων

α	num. of iterations
.5	34
.75	81
.8	104
.85	142
.9	219
.95	449
.99	2292
.999	23015

Ο πίνακας υπερσυνδέσμων **H**

- Διάφορες προσαρμογές μπορεί να γίνουν πάνω στον H
- Στην βασική υλοποίηση, κάθε εξερχόμενος σύνδεσμος έχει το ίδιο "βάρος/σημαντικότητα"
- Παρόλο που η τακτική αυτή είναι δημοκρατική, εύκολη στην υλοποίηση, εντούτοις δεν είναι η κατάλληλη για τα rankings
- Στην πραγματικότητα, ο random surfer δεν διαλέγει τυχαία με την ίδια πιθανότητα ποιον σύνδεσμο θα ακολουθήσει, αλλά λαμβάνει υπόψη του το πλούσιο περιεχόμενο των σελίδων όπου θα πάει, αλλά και το κείμενο πάνω στους υπερσυνδέσμους
- Έτσι, αντί για την υπόθεση του random surfer, έχουμε τον **intelligent surfer**

Παράδειγμα προσαρμοσμένου πίνακα **H**

- Πώς αποφασίζουμε με ποιο τρόπο θα αναθέσουμε διαφορετικά βάρη στους εξερχόμενους υπερσυνδέσμους;
- Από τα access logs!
- Παράδειγμα: Από την P_1 είναι δυο φορές πιο πιθανό να πάμε στην P_2 παρά στην P_3
- Προφανώς όλες οι παρόμοιες μέθοδοι θα είναι ευρεστικές
- Για παράδειγμα, τα στοιχεία $\mathbf{H_{45}}$ και $\mathbf{H_{46}}$ μπορούν να προσδιοριστούν με βάση την ομοιότητα (cosine similarity) μεταξύ των σελίδων P_4 με την P_5 και P_6
- Για το γράφημα με τους 6 κόμβους ο νέος πίνακας H θα μετατραπεί στον ακόλουθο:

Παράδειγμα προσαρμοσμένου πίνακα **H**

$$\mathbf{H} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$\mathbf{H}' = \begin{pmatrix} 0 & 2/3 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Ο πίνακας τηλεμεταφοράς **E** (1/3)

- Μια από τις πρώτες προσαρμογές ήταν ότι αντί για τη χρήση του $1/n\mathbf{e}\mathbf{e}^T$ προτιμήθηκε ο πίνακας $\mathbf{e}\mathbf{v}^T$
- Το $\mathbf{v}^T$ με $\mathbf{v}^T > 0$, είναι ένα διάνυσμα πιθανοτήτων που ονομάζεται **personalization** ή **teleportation διάνυσμα**
- Αφού το $\mathbf{v}^T$ είναι διάνυσμα πιθανοτήτων με θετικά στοιχεία, κάθε κόμβος είναι συνδεδεμένος με κάθε άλλο κόμβο, άρα ο **G** είναι πρωτογενής
- Χρησιμοποιώντας το $\mathbf{v}^T$ αντί για το $1/n\mathbf{e}^T$ σημαίνει ότι οι πιθανότητες τηλεμεταφοράς δεν είναι πλέον ομοιόμορφες

Ο πίνακας τηλεμεταφοράς **E** (2/3)

- Άρα για κάθε τηλεμεταφορά, ο surfer δεν επιλέγει ομοιόμορφα σε ποια σελίδα θα πάει, αλλά καθοδηγείται από το διάνυσμα $\mathbf{v}^T$
- Αυτή η μετατροπή ευτυχώς δεν καταστρέφει τα πλεονεκτήματα της power method
- Όταν $\mathbf{G}=\alpha\mathbf{S}+(1-\alpha)\mathbf{e}\mathbf{v}^T$, τότε η power method γίνεται:

$$\begin{aligned}\pi^{(k+1)T} &= \pi^{(k)T}\mathbf{G} \\ &= \alpha\pi^{(k)T}\mathbf{S} + (1-\alpha)\pi^{(k)T}\mathbf{e}\mathbf{v}^T \\ &= \alpha\pi^{(k)T}\mathbf{H} + (\alpha\pi^{(k)T}\mathbf{a} + 1 - \alpha)\mathbf{v}^T\end{aligned}$$

Ο πίνακας τηλεμεταφοράς **E** (3/3)

- Αυτή η αλλαγή δεν έχει καμία επίδραση πάνω
  - στο ρυθμό σύγκλισης
  - στον πολλαπλασιασμό διανύσματος με αραιό πίνακα
  - στις μικρές αποθηκευτικές απαιτήσεις
- Όμως, αλλάζει το ίδιο το διάνυσμα PageRank!!
- Αυτό δεν είναι μειονέκτημα !?
- Δεν είναι απαραίτητο ότι σε όλους μας "ταιριάζει"το ίδιο ranking
- Άλλωστε, παρέχει μια ευελιξία ώστε ανάλογα τις ανάγκες μας να προσαρμόζουμε απλά το $\mathbf{v}^T$

Προσωποποίηση του PageRank

- Η προσωποποίηση αλλάζει το διάνυσμα PageRank, από query-independet και user-independent σε user-dependent και πιο δύσκολο στον υπολογισμό
- Στην θεωρία είναι ωραία η προσωποποίηση, αλλά στην πράξη είναι δύσκολα εφαρμόσιμη
  - Κάθε $\boldsymbol{\pi}^T$ απαιτεί μερικές ημέρες για τον υπολογισμό του
- Οπότε, αφού επικρατεί η άποψη ότι η προσωποποιημένη αναζήτηση είναι η μελλοντική τάση στις μηχανές αναζήτησης, αρκετοί δημιούργησαν ψευδο-προσωποποιημένα διανύσματα PageRank
  - Δεν απευθύνονται σε κάθε χρήστη, αλλά σε ομάδες χρηστών

Topic-sensitive PageRank (1/3)

- Δημιουργία ενός πεπερασμένου αριθμού PageRank διανυσμάτων $\boldsymbol{\pi}^T(\mathbf{v}_i^T)$, κάθε ένα από αυτά πολωμένο ως προς κάποια συγκεκριμένο θέμα
- Ποια θέματα επιλέχθηκαν;
- Ο Taher Haveliwala επέλεξε τα 16 πρώτα από το Open Directory Project (ODP)
- Τα 16 πολωμένα διανύσματα προϋπολογίζονται
- Το ζήτημα είναι να τα συνδυάσουμε αποτελεσματικά κατά την ερώτηση του χρήστη

Topic-sensitive PageRank (2/3)

- Ο Taher Haveliwala έφτιαξε έναν κυρτό συνδυασμό αυτών ως εξής

$$\boldsymbol{\pi}^T = \beta_1 \boldsymbol{\pi}^T(\mathbf{v}_1^T) + \beta_2 \boldsymbol{\pi}^T(\mathbf{v}_2^T) + \ldots + \beta_{16} \boldsymbol{\pi}^T(\mathbf{v}_{16}^T)$$

όπου $\Sigma\beta_i = 1$

- Για παράδειγμα, η ερώτηση *science project ideas* εμπίπτει μεταξύ των εξής κατηγοριών του ODP:
  - Κατηγορία 7: Kids και Teens
  - Κατηγορία 10: Reference
  - Κατηγορία 12: Science
- Προφανώς τα αντίστοιχα διανύσματα αυτών των κατηγοριών πρέπει να πάρουν μεγαλύτερο βάρος ή ίσως και όλο το βάρος

Topic-sensitive PageRank (3/3)

- Για τον υπολογισμό των βαρών χρησιμοποιήθηκε ένας classifier Bayes
- Όταν υπολογιστεί το topic-sensitive score, συνδυάζεται με το αντίστοιχο content score

- Ο Jeh Glen, Taher Haveliwala & Serendap Kamvar δημιούργησαν το καλοκαίρι του 2003 την εταιρεία Kaltix για να προωθήσουν την ιδέα του personalized PageRank, και τελικά η εταιρεία τους αγοράστηκε το Σεπτέμβριο του 2003 από την Google
- Τον Μάρτιο του 2004, η Google προώθησε την προσωποποίηση http://labs.google.com/personalized

Το φάσμα του personalized πίνακα **G** (1/4)

- ΘΕΩΡΗΜΑ: Εάν το φάσμα (ιδιοτιμές) του στοχαστικού πίνακα **S** είναι $\{1,\lambda_2,\lambda_3,\ldots,\lambda_n\}$, τότε το φάσμα του personalized πίνακα Google $\mathbf{G}=\alpha\mathbf{S} + (1-\alpha)\mathbf{e}\mathbf{v}^T$ είναι $\{1,\alpha\lambda_2,\alpha\lambda_3,\ldots,\alpha\lambda_n\}$, όπου το $\mathbf{v}^T$ είναι ένα διάνυσμα πιθανοτήτων

Το φάσμα του personalized πίνακα **G** (2/4)

- Αφού ο **S** είναι στοχαστικός, τότε το (1,**e**) είναι ένα ζεύγος του **S**
- Έστω ότι **Q** = (**e** **X**) είναι μη ιδιόμορφος (non-singular) πίνακας που έχει το ιδιοδιάνυσμα **e** ως πρώτη στήλη του
- Έστω ότι $\mathbf{Q}^{-1} = \begin{pmatrix} \mathbf{y}^T \\ \mathbf{Y}^T \end{pmatrix}$
- Τότε $\mathbf{Q}^{-1}\mathbf{Q} = \begin{pmatrix} \mathbf{y}^T\mathbf{e} & \mathbf{y}^T\mathbf{X} \\ \mathbf{Y}^T\mathbf{e} & \mathbf{Y}^T\mathbf{X} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$
- Απ' εδώ παίρνουμε δυο χρήσιμες ταυτότητες
  - $\mathbf{y}^T\mathbf{e}=1$
  - $\mathbf{Y}^T\mathbf{e}=\mathbf{0}$

Το φάσμα του personalized πίνακα **G** (3/4)

- Ως συνέπεια, ο μετασχηματισμός ομοιότητας

$$\mathbf{Q}^{-1}\mathbf{S}\mathbf{Q} = \begin{pmatrix} \mathbf{y}^T\mathbf{e} & \mathbf{y}^T\mathbf{S}\mathbf{X} \\ \mathbf{Y}^T\mathbf{e} & \mathbf{Y}^T\mathbf{S}\mathbf{X} \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{y}^T\mathbf{S}\mathbf{X} \\ 0 & \mathbf{Y}^T\mathbf{S}\mathbf{X} \end{pmatrix}$$

φανερώνει ότι ο $\mathbf{Y}^T\mathbf{S}\mathbf{X}$ περιέχει τις υπόλοιπες ιδιοτιμές του **S**, $\lambda_2,\lambda_3,\ldots,\lambda_n$

Το φάσμα του personalized πίνακα **G** (4/4)

- Εφαρμόζοντας τον μετασχηματισμό ομοιότητας στον $\mathbf{G}=\alpha\mathbf{S} + (1-\alpha)\mathbf{e}\mathbf{v}^T$

$$\begin{aligned}
\mathbf{Q}^{-1}(\alpha\mathbf{S} + (1-\alpha)\mathbf{e}\mathbf{v}^T)\mathbf{Q} &= \alpha\mathbf{Q}^{-1}\mathbf{S}\mathbf{Q} + (1-\alpha)\mathbf{Q}^{-1}\mathbf{e}\mathbf{v}^T\mathbf{Q} \\
&= \begin{pmatrix} \alpha & \alpha\mathbf{y}^T\mathbf{S}\mathbf{X} \\ 0 & \alpha\mathbf{Y}^T\mathbf{S}\mathbf{X} \end{pmatrix} + (1-\alpha)\begin{pmatrix} \mathbf{y}^T\mathbf{e} \\ \mathbf{Y}^T\mathbf{e} \end{pmatrix}\begin{pmatrix} \mathbf{v}^T\mathbf{e} & \mathbf{v}^T\mathbf{X} \end{pmatrix} \\
&= \begin{pmatrix} \alpha & \alpha\mathbf{y}^T\mathbf{S}\mathbf{X} \\ 0 & \alpha\mathbf{Y}^T\mathbf{S}\mathbf{X} \end{pmatrix} + \begin{pmatrix} (1-\alpha) & (1-\alpha)\mathbf{v}^T\mathbf{X} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\
&= \begin{pmatrix} 1 & \alpha\mathbf{y}^T\mathbf{S}\mathbf{X} + (1-\alpha)\mathbf{v}^T\mathbf{X} \\ \mathbf{0} & \alpha\mathbf{Y}^T\mathbf{S}\mathbf{X} \end{pmatrix}
\end{aligned}$$

- Επομένως, οι ιδιοτιμές του $\mathbf{G}=\alpha\mathbf{S} + (1-\alpha)\mathbf{e}\mathbf{v}^T$ είναι οι $\{1,\alpha\lambda_2,\alpha\lambda_3,\ldots,\alpha\lambda_n\}$